情報アクセス対話のための質問応答技術評価タスク

加藤　恒昭[†]・福本　淳一[††]・桝井　文人[†††]・神門　典子[††††]

あるトピックに関して対話的に行われる一連の情報アクセスを質問応答システムが支援する能力，情報アクセス対話の対話相手として情報を提供するために質問応答システムが持つべき能力を定量的に評価するためのタスクを提案する．このタスクでは，対話の実現の基本となる対話文脈を考慮した質問の解釈，つまり照応解消や省略処理等のいわゆる文脈処理の能力を評価する．本稿では，タスクの設計を示し，その根拠となる調査結果を報告する．提案するタスクは以下の点で新規かつ有益である．対話的情報アクセスを対象として，そこで必要な質問応答技術が効果的に評価できるという課題設定と構成の独自性を持つ．評価尺度については応答の自然性において問題となる回答の質や回答列挙の体系の違いに配慮し，複数の体系を許す多段階評価手法を備えている．システムの文脈処理能力をある程度まで切り離して評価することを可能とする参照用テストセットと呼ぶ枠組みを有している．

キーワード：質問応答，評価タスク，対話，文脈処理

Evaluation Task of Question Answering for Information Access Dialogues

Tsuneaki Kato[†], Jun'ichi Fukumoto[††], Fumito Masui[†††] and Noriko Kando[††††]

A novel task for evaluating question answering technologies is proposed. This task assumes interactive use of question answering systems and evaluates among other things, the abilities needed under such circumstances, i.e. proper interpretation of questions under a given dialogue context; in other words, context processing abilities such as anaphora resolution and ellipses handling. This paper shows the design of the task and its empirical background. The task proposed is not only novel as an evaluation of the handling of information access dialogues, but also includes several valuable ideas such as a measuring metric in order to obtain intuitive evaluation of the answers to list-type questions and reference test sets for obtaining information on context processing ability in isolation.

Key Words: *Question Answering, Evaluation Task, Dialogue, Context Processing*

[†] 東京大学, The University of Tokyo
[††] 立命館大学, Ritsumeikan University
[†††] 三重大学, Mie University
[††††] 国立情報学研究所, National Institute of Informatics

1 はじめに

質問応答技術は自然言語によって表現された質問に文書でなく情報そのもので回答する事を可能とするもので，情報アクセスの新しい形として期待されている (Voorhees and Tice 2000). 事実に関する独立した質問に一問一答形式で回答するものを中心に研究が始められたが，近年は様々な面で研究の展開が見られ，そのひとつに対話性の重視があげられる.

質問応答技術を牽引してきたといってよいTREC (Voorhees 2005; NIST 2007) では，TREC2001において対話的な利用を前提とした文脈処理の能力を評価する試みがなされている (Voorhees 2001). その後，TREC 2004から，相互に独立した質問ではなく，あるトピックに関する一連の質問の集まりという形で課題を与えるようになっている (Voorhees 2004). 文脈処理の能力を評価するものでないとはいえ，あるトピックに関して一連の質問を行うという利用場面が自然であると考えられている点が注目される. また，あるトピックに関する複数の質問にどの程度回答できるかを，複数文書要約の評価指標とすることが試みられており (Mani, House, et al. 1998)，ここでも，あるトピックに関する一連の質問に回答できることが重視されている.

一連の質問に回答するという利用形態は質問応答システムの進むべき方向のひとつとしても議論されており，例えば，新人レポータがある事件の記事を執筆するために，彼の記事で答えられるべき大きな質問をより簡単な質問の集まりに言い換えてシステムに訊ねるという形で，アナリストやレポータが利用しうる質問応答システムへの発展が提案されている (Burger, Cardie, et al. 2001). また，ARDA の AQUAINT program (ARDA 2007) ではアナリストが分析的に用いる質問応答システムの構築がその目的とされており，より積極的に対話的な質問応答の研究が進められている. 質問の分解を含めて，分析的説明的な質問にどう答えるか，明確化等の利用者とのやりとりはどうするか等が研究の関心となっている (Hickl, Lehmann, Williams, and Harabagiu 2004; Small, Shimizu, et al. 2003).

本稿では，あるトピックに関して対話的に行われる一連の情報アクセスを質問応答システムが支援する能力，情報アクセス対話の対話相手として情報を提供するために質問応答システムが持つべき能力を定量的に評価するためのタスク，IAD タスク[1]を提案する. 質問応答システムが情報アクセス対話に参加するために必要となる様々な能力 (Burger et al. 2001) の中で，IAD タスクでは，そもそも情報アクセス対話を扱うためにはどのような質問に答えられる必要があるのか，そして，対話の実現の基本となる対話文脈を考慮した質問の解釈，つまり照応解消や省略処理等のいわゆる文脈処理はどの程度必要なのかに着目し，その能力を評価する.

IAD タスクは，情報アクセス技術に関する一連の評価ワークショップ NTCIR Workshop (NII 2007) において，NTCIR-4 の QAC2 Subtask 3(Kato, Fukumoto, and Masui 2004b; Kato, Fuku-

[1] IAD は情報アクセス対話 (Information Access Dialogue) の頭文字からとった.

moto, Masui, and Kando 2005)，NTCIR-5 の QAC3(Kato, Fukumoto, and Masui 2004a; 加藤，福本，桝井，神門 2006) として実施されたものに基づいている．対話的な質問応答というそもそものアイディアは NTCIR-3 の QAC1 Subtask 3(Fukumoto, Kato, and Masui 2003) に遡るが，NTCIR-4 の QAC2 Subtask 3 での実施においてタスクの抜本的な改変を行い本稿で述べる形態を固め，同時にタスクの裏付けについての実験を行った．その後，そこでの経験を基に幾つかの洗練を行って，NTCIR-5 の QAC3 として実施している．

ここで，評価タスクの提案という本稿の特殊性について一言述べておく．研究や技術の進展や加速のために共通の評価が必要であり．それを得るための評価タスクが重要であることは，議論の余地がまったくないとはいえないまでも (関根，影浦，奥村，乾 2005)，大概の合意を得ていると思われる (小川，佐々木，増山，村田，吉岡 2002)．一方で個々の評価タスクについて考えると，ある評価タスクが価値あるものであるためには，それが評価する研究や技術が評価されるに値するものであり，かつ，その評価のために適切に設計されている必要がある．前者は研究や技術の価値の議論であり，後者も何をもって適切とするかが絡んで必ずしも明快な議論とはならない．本稿では，ここで提案する IAD タスクにおいて高い評価を得たシステムあるいは技術が可能とする利用場面を示し，前者の根拠とする．加えて，後者については，少なくとも 2 回の実施を通じて明らかとなった問題について一定を解決を与えていることを根拠とする．設計ということで一部に恣意的な決定を含んでいるし，この評価タスクであらゆるデータが収集できるわけではない．実施できなければならないという現実性との妥協もある．そのような一連の留保を前提にしているとはいえ，本提案が，課題設定の独自性，評価に関する様々な配慮，情報収集のための仕組み等の点で，新規かつ有益なものであることを主張する．

本稿の構成は以下の通り．2 節で IAD タスクの枠組みを説明する．タスク設計の中心となる質問シリーズを説明し，それがトピック推移の観点から収集型とブラウジング型に分類されることを述べる．加えて，IAD タスクの枠組みの根拠となった実験結果を示し，このタスクによって評価される技術が可能とする質問応答技術の利用場面を示唆する．3 節では，評価の枠組みとして，回答の列挙に複数の体系を許し回答の 2 種類の質を考慮した多段階評価手法を提案する．そして，なぜそのような枠組みが必要であるかを実例に基づいて説明する．4 節ではより多くの情報を得るための補助的な仕組みとしての参照用テストセットについて説明し，それがシステムの文脈処理能力をある程度まで切り離した評価を可能とすることを示す．5 節では関連する取り組みを述べ，それとの比較を通じて本提案の有効性を示し，特に収集型とブラウジング型への分類を含む質問シリーズの構成方法が重要であることを述べる．6 節で全体をまとめる．また，IAD タスクに対して，最先端のシステムがどのような結果を示すのかを付録にまとめた．

2 タスクの枠組み

IADタスクは，対話的な情報アクセスでの質問応答システムの利用を考え，そこで必要な照応解消や省略処理等のいわゆる文脈処理の能力を評価することを目的とする．様々なバラエティを持つ情報アクセス対話の中で，特に，与えられたトピックについてのレポートを書くための素材となる情報を得るような対話を想定している．これはある事件の記事を執筆するために，必要な情報を比較的簡単な質問の集まりとしてシステムに訊ねるという形態とも近い．

2.1 質問シリーズ

IADタスクでは，システムに一連の質問（シリーズと呼ぶ）を与え，それに次々と回答させてゆく．シリーズの先頭以外の質問は，それ以前の質問の一部もしくはその回答を参照する照応表現を含んでいる[2]．この一連の質問とそれへの回答が情報アクセス対話を構成する．実際の利用場面ではシステムは対話的に質問に回答することが期待されるが，本タスクではその対話性は模擬されるだけで，複数のシリーズ（テストセットと呼ぶ）をバッチ的に与え，それに回答することをシステムに求める．ここで，システムはある質問がシリーズの先頭であるという情報は利用してよいが，ある質問に回答する際にそれに続く質問を参照することは許されない．これは本タスクが対話的な状況でのシステムの利用を模擬していることからの制約である．対話の展開があらかじめ定められていることは対話本来のダイナミクスを失わせているが，その一方で，本タスクの実施に参加したシステムがすべて同じ質問に回答するので，相互比較可能な結果が得られることに加え，正解をプーリングすることでテストセットが再利用可能となるという利点を有している．

IADタスクでは大きく分けて収集型とブラウジング型という2種類のシリーズを設定している．これは，情報アクセス対話が，利用者があるトピックについてのレポートや要約を作成するための情報を収集する等の目的でそれに関する一連の質問を行なうような対話（収集型）と，利用者の興味の赴くところに従って対話の進行と共にトピックが変わっていくような対話（ブラウジング型）との2つの極を持つという直観に基づいている．タスクにおいて，あるシリーズがどちらの型に属するかは与えられず，システムはそれを自分で判定しなければならない．IADタスクが想定している情報アクセス対話は与えられたトピックについての様々な情報を収集するもので，当然収集型の対話が支配的であるが，後述するように実際の場面ではその部分部分にブラウジング的な要素が含まれる．これが本タスクにブラウジング型を含め，かつシリーズの型の同定をシステムに求めている理由である．なお，シリーズ単位で型を区別したことには分析が容易になることへの期待がある．

[2] 省略やゼロ代名詞，英語の定名詞句に相当する一般名詞の反復を含む．表層から明らかでないので不適切かもしれないが，「表現」と呼ぶことにする．

IADタスクのシリーズの例を図1に示す．収集型は，広い意味で共通のトピックに関する質問からなり，そのトピックはシリーズ先頭の質問で導入される．すべての照応表現がそのトピックを参照するというのがもっとも厳しい意味での収集型（狭義の収集型と呼ぶ）である．図1のSeries 2-14はそのような収集型で，先頭質問で述べられている「小沢征爾」を補うことで，すべての質問の文脈処理が行える．一般には，複数の照応表現を持ち，その一方がトピックを参照するような質問や，トピックが関連した出来事やその一般化を参照するような表現をもつ質問等も収集型のシリーズに含まれる．Series 2-20はその例で．第3問は複数の照応表現を含み，第6問は先頭質問文で述べられているトピックであるジョージ・マロリーが関連したイベントを参照している．ブラウジング型はそのような大域的なトピックを持たず，質問中の照応表現は，直前の質問の回答や以前の質問中で言及された事物を参照している．Series 2-22はブラウジング型の例である．

Series 2-14
小沢征爾さんはいつ生まれましたか。
どこの生まれですか。
大学はどこを卒業しましたか。
師事した先生は誰でしたか。
誰に認められましたか。
98年にはどこで指揮を行っていましたか。
2002年からどこの指揮者になりますか。
Series 2-20
ジョージ・マロリーはどこの国で生まれましたか。
彼の有名な言葉は何ですか。
それを言ったのはいつのことですか。
彼が初めて山に登ったのは何歳の時ですか。
彼がエベレストの頂上付近で行方を絶ったのは何次遠征のときですか。
それは何年のことですか。
彼が最後に目撃されたのはエベレストの何メートル付近ですか。
彼の遺体を発見したのは誰ですか。
Series 2-22
ニューヨーク・ヤンキースの本拠地となっている球場はどこですか。
何年に造られたものですか。
そこには何人の記念碑が飾られていますか。
1999年に飾られたのは誰ですか。
彼が新婚旅行で来日したのは何年ですか。
その時の結婚相手は誰ですか。
彼女をポップ・アートに描いているのは誰ですか。
彼が描く缶詰はどこの会社のものですか。

図1　シリーズの例

2.2 個々の質問の範囲

IADタスクのシリーズを構成する質問は，疑問代名詞を含む文の形式を持ち，名称を正解とする質問である．ここで，名称というのは，人名や組織名等いわゆる固有表現に留まらず，日付け，数量を含み，種の名称，機械や身体的部分等の一般名称を含む．統語的には複合名詞が正解の範囲とほぼ重なるが，小説や映画のタイトル等そこから外れるものも含まれる．システムはこれらの名称をそれを含んだ部分でなく，過不足なく抜き出してひとつの回答とし，複数の正解があると判断される場合はそれらをリストとして列挙することを求められる．質問の正解が知識源中に存在することは保証されていないので，回答が存在しないこと，空リストが正解ということもありうる．各回答（回答リストの要素）は，それを抜き出した文書であり，それが正解であることの根拠となる文書の識別子を伴っていなければならい．

回答リストの要素として日付や数量を含む名称の表現を過不足なく抜き出すことを要求すること，回答リストとしてすべての回答の過不足ない列挙を求めることは，文書でなく情報そのもので回答するという質問応答の流れから当然と考えるが，実際にタスクとして実施する場合，細部の検討が必要となる．日付や数量の表現については，質問への自然な回答を可能とするため，以下の表現も正解範囲であることを明示する必要がある．なお，名称という正解範囲の根拠づけは2.3節において，過不足のない列挙の問題は評価に関する3節において論じる．

数値表現に属性の詳細化具体化を行うための表現が付属したもの 「年間300台」「タテ50 cm ヨコ30 cm」「一人当たり3リットル」「重さ3トン」等．

範囲表現（定型的，慣用的なもの） 「10～12%」「8世紀後期から9世紀初期」「四国から九州まで」「30人以上」「30人以上50人以下」等．「東京大阪間」「羽田—千歳」「千葉県内」等，空間的な範囲表現（区間表現）も含む．

概数表現（蓋然表現） 「約100人」「3億円程度」等．「シカゴ近郊」「東京都近辺」「舞浜駅前」「大使館裏」等，空間的な蓋然表現も含む．

これらを正解の範囲としない場合，まず，「どのくらい利用されていますか」に年間なのか月間なのかが不明確であるような「300台」と回答する，「どのくらいの大きさですか」に「50 cm」と長さで回答する等の不自然さを強いることになる．不自然さの問題に加え，これらを許さないことは正解の網羅的な列挙や重複の判断でも問題となる．「タテ50 cm ヨコ30 cm」と回答できずに「50 cm」「30 cm」の両方を挙げる必要があるとか，「10～12%」において，「12%」は抜き出しという形で得られるが，「10」だけでは単位が含まれないので正解として抜き出せないとか，「約100人」は「102人」と同一の情報としていいかもしれないが「100人」はどうか等の問題が生じてくる．

2.3 タスクの根拠

IAD タスクの根拠として，レポート作成の情報を得るための対話的情報アクセスで名称を正解の範囲とする質問応答システムが使われうるのか，そして，その状況での質問にはどのような照応表現がどの程度含まれるかを調査した[3]．

2.3.1 データ収集

調査は，IAD タスクが前提とする状況で利用者から発せられるであろう質問を収集し，分析することで行った．新聞記事から選択した人物，組織，出来事等のトピックを被験者に提示し，それに関するレポートを執筆するという状況を設定した．レポートは与えられたトピックの事実関係をまとめたもので予測や意見はそこに含めないものとし，質問の文型は疑問代名詞を含む Wh 型に限定するように指示した．以下の 2 種類の収集を実施した．

アンケート方式による調査　レポートに含めたいと考える情報を質問文の形式で表現するように指示することで，レポート執筆のための一連の質問を作成させた．作成する質問数は 1 トピックあたり 10 問を目安とした．作成した質問に次々と回答が得られるという想定で，ひとつのトピックについて複数の質問を作成させ，質問中に代名詞等の表現を含めることを許した．これにより自然な質問の系列が作成されることを期待した．60 のトピックについて，30 人の被験者に一人あたり 30 トピックを割り当てた．トピックの提示は 20 文字程度の短い記述，それについての短い記事，それについての記事 5 編，と 3 種類の方法を均等に混ぜた．集めたデータのうち，40 トピックについての各 9 系列を構成する Wh 質問，3,401 文を分析した[4]．

WOZ 方式による調査　レポート執筆という状況設定で，質問を事前に考えたのち，WOZ 方式で模擬された質問応答システムと情報アクセス対話を行うことで情報収集を行わせた．質問数は 1 トピックあたり 10 問を目安とした．20 のトピックについて，6 人の被験者に各 10 トピックを割り当て，被験者にはトピックと 100 文字程度の概要を提示した．WOZ 役の協力者は 4 名で，事前に担当するトピックについて 800 文字から 1,600 文字程度の要旨を作成するという事前準備をしており，作成した要旨，新聞記事全文検索システム，自分の記憶を用いて，利用者からの質問に対話的に回答した．対話はキーボードを用いて行った．被験者には事実に関する簡単な質問に回答できる質問応答システムを利用していると説明し，WOZ 役にも理由や意見を訊ねる質問については回答できないと応答する，必要な場合は問い返しを行ってかまわない，回答は簡潔を旨とするが自然な協調的振る舞いを禁じるものではない等，その役割を教示した．集めたデータすべて，20 ト

[3] ここで用いたデータ収集の手法はテストセット構築にも利用できる．なお，これらの調査は本提案の基となった NTCIR-4,5 での実施におけるテストセット構築と並行して行ったものである．
[4] トピックの提示方法の詳細，分析データ選択の過程等については (加藤，福本，桝井，神門 2004b) に詳しい．

ピックについての各3系列を構成する質問等，620文を分析した[5]．

2.3.2 質問と回答のタイプに関する分析

質問の種類，質問が何をたずねているかを分類した結果を表1に示す．ここで，4W 質問は「小沢征爾氏は誰に師事しましたか」のように具体的な人名等を訊ねる質問で，「～って誰ですか」「～とは何ですか」という質問は定義・説明・記述を訊ねる質問に分類している．WOZ 方式の収集において，YesNo 質問の場合はそれに対する協調的応答の内容から判断して訊ねている内容を決定した．

表2は回答のタイプによる分類である．ここで，「一般名称」は名称から数量や日付の表現，固有表現（固有名称）を除いたものである．「固有名称」には小説や映画のタイトルが含まれる．この分類は表1に示した分類と強く関連する．例えば Why 質問に回答するためには一般に節や文が必要となる．しかし一方で，4W 質問に分類された質問がすべて名称によって回答できるわけではない．例えば，場所を訊ねる質問でも「ロブスタが好んで住むのはどこですか」には名称での回答は困難で，一定の量の記述や説明を必要とする．アンケート方式では，この分析を質問だけを見ることで行ったため，幾つかの質問については確定的な分類が行えなかった．「たぶん名称」と分類されたものは「AIBO の由来は何ですか」のような質問で，AIBO が何かのアクロニムであれば名称の範囲に収まるが，その由来が長い物語となるかもしれないものである．このように質問だけからは予想される回答が複数のカテゴリにまたがるものは他の

表 1 質問で訊ねている内容の分類

何を訊ねているか	アンケート方式	WOZ 方式
4W 質問（誰，何，どこ，いつ等．数量に関する質問を含む）	70.4%	67.1%
Why 質問（なぜ等，理由を訊ねるもの）	4.4%	6.5%
How 質問（どうやって等，手順や手法をを訊ねるもの）	9.8%	17.0%
定義・説明・記述を訊ねる質問	15.5%	10.8%

表 2 （推測される）回答による分類

何で答えられるか	アンケート方式	WOZ 方式
数量や日付表現	27.8%	31.0%
固有名称	20.6%	22.1%
一般名称	9.5%	8.8%
たぶん名称	16.7%	—
節・文・文章	25.3%	38.1%

[5] 13%程度の YesNo 質問や命令文が含まれている．それらの扱いを含めて，ここで論じていない明確化発話や協調的応答の分析については (Kato, Fukumoto, Masui, and Kando 2006) に詳しい．

分類の間でも存在するが，簡単のためにそれらは複雑な方に分類した．WOZ 方式の場合，分類は WOZ 役の発話に基づいて行ったが，発話全体の形式ではなく．質問への回答そのものに注目した．例えば，「いつ生まれましたか」への回答の「3 月 13 日に生まれました」の場合，その分類は節や文ではなく日付表現である．

2.3.3 照応表現の特徴に関する分析

質問中に含まれる照応表現として，前方照応のための手段を，指示代名詞（連体詞を含む），ゼロ代名詞，英語等の定名詞句に相当する前出名詞の繰り返し，省略の 4 つに分けて，その出現頻度を調べた．2 つの状況を比較するために（出現頻度 /（質問文数－系列先頭の質問文数））で計算される相対頻度をまとめたものを表 3 に示す．合計は 100%を越えるが，例えば，「**それまで誰が**その国の指導者だったのですか」のように複数の照応表現がひとつの質問文中に含まれる場合があるためである．

省略を除く照応表現のうち，与えられたトピック，つまり大域的トピック以外を参照するものの割合は，アンケート方式で 29%，WOZ 方式で 22%であり，そのうち同じ質問文中に大域的トピックを参照する表現を持たないものがそれぞれ 92%, 81%であった．このような質問の存在は質問系列中で焦点が推移しており，大域的トピックでないものが焦点となっていることを示している．

系列の先頭以外で，照応表現を含まない質問のうち，アンケート方式で 55%，WOZ 方式で 68%が，焦点となっているものを代名詞化しないでそのまま表現するケースであった．これは例えば，人物を姓のみで参照する場合や，ニホンカワウソやハイブリッド車等．名詞で表現されるクラスが焦点となっている場合で，前出名詞の繰り返しとも考えられるものである．それ以外は焦点の変化と関連する．例えば，あるニュース番組のダイオキシン汚染に関する誤報道をトピックとした場合に，その番組に対する一連の質問に続いて「ダイオキシンの毒性はどのくらいですか」と訊ねるような場合，逆にチャールズ皇太子が与えられたトピックで，その息子達に関する質問が幾つか続いた後に，「チャールズ皇太子の長年の恋人とは誰ですか」と焦点が戻る場合等がある．WOZ 方式では質問の回答に含まれた内容へと焦点が移る場合もあった．

表 3　質問中に表れる照応表現

分類	アンケート方式	WOZ 方式
照応表現を含まない質問	30.0%	25.5%
代名詞	15.7%	2.5%
ゼロ代名詞	45.6%	56.6%
定名詞句相当	9.8%	18.6%
省略	0.5%	0.2%

2.3.4 考察

表2からわかるように，レポート作成のための質問のうち，アンケート方式で58%–75%，WOZ方式で62%が数量等を含む名称を回答とする質問となる．レポートを執筆するための情報を訊ねる質問を収集した状況では，節や文で回答することが多いと思われる「なぜ」を訊ねる質問は少なく，説明や定義を求める質問も予想されたほど多くはない．これは，「小沢征爾って誰ですか」という質問が，例えば彼の誕生日や出身地を訊ねるような具体的な質問に展開されているためであると考えられる．60%強という数字は決定的ではないが，名称を正解の範囲とするような質問応答システムはこのような状況で充分に有用であると判断できる．ちなみに，アンケート方式で収集した質問のうち，名称を回答とする737問について，その正解が新聞記事集合から得られるかを調査したところ，84%について正解が得られ，新聞記事等の大規模文書を知識源とすることが現実的であることもわかる．

更に重要なことは，このような状況で得られた質問文に様々な照応表現が含まれることである．照応表現が頻出することに加えて，その参照先は単に情報収集の中心となる大域的なトピックに限られるような簡単なものではない．情報アクセス対話は，その焦点が対話の進行によって推移し，サブダイアログも含む複雑なものともなりうるため，それに応じた文脈処理が必要であることがわかる．

ここで示された状況が，IADタスクの設定で模擬されている．IADタスクが評価するのは，ここで示された状況に対応し，対話的な情報アクセスを実現するための質問応答技術であり，このタスクで高い評価を得た技術は，本節の実験で模擬されたような対話的な情報アクセスの実現に有効である．

3 評価手法

3.1 対話性に伴う問題

IADタスクでは，各質問に対して，存在しないことを含めていくつ存在するかわからない正解を過不足なく収集し，それらすべてを列挙したリストをひとつ返すことを求める．正解数は問題毎に異なり事前に与えられないので，個々の質問に関する評価は精度と再現率の両方を考慮したF値を採用する．システムの総合評価はその評価のテストセット全体の平均である．情報検索一般とは異なる質問応答の特殊性から普通のF値ではなく，様々な配慮が必要となるが，これについては3.2節で述べる．ある回答が正解であるかは，回答とそれと合わせて提示される根拠記事の適切性によって判断される．質問と無関係な記事を根拠としていれば文字列として正解と同一であっても不正解として扱われる．

対話的な情報アクセスという特徴から質問の解釈が文脈に依存し，それが正解に影響するという問題がある．IADタスクでは，質問の正解は判定者である人間が適切と判断した文脈の下

でおこなった解釈によって決定され，システムの解釈やシステムのそれ以前の回答とは無関係であるとする．例えば，図1のSeries 2-22の2番目の質問の正解は，常にニューヨーク・ヤンキースの本拠地であるヤンキースタジアムが建てられた1923年であり，システムが最初の質問にシェイスタジアムと誤って答え，2番目の質問にそれが建てられた年である1964年を“正しく”回答しても不正解とする．一方，最初の質問にシェイスタジアムと答えていても，適切な根拠記事と共に1923年を回答していれば，2番目の質問については正解と判断される．特に後者については若干の違和感があるが，システムが文脈を内包的に管理し，2番目の質問を「ニューヨーク・ヤンキースの本拠地となっている球場は何年に造られたものですか」と解釈したと考えれば，不正解にする理由はない．また，収集型のシリーズでは，質問文は直前の質問や回答よりもシリーズの先頭で導入されたトピックを参照していることが多く，直前の質問に正解することが現在の質問に正解する必要条件になっている場合は必ずしも多くない．これらの理由に加えて，システムが起こしうる誤った解釈すべてについてその後の正解がどうあるべきかを事前に判断するのは不可能ということから，このような方式としている．

3.2 評価尺度

対話性の問題以外に，可能な正解すべてを列挙したリストをひとつ返すことを求めるリスト型課題の評価には以下のような難しさがある (加藤，福本，桝井，神門 2004a)．

重複の扱い　同じ事物を指す複数の表現，人名における役職の有無，外人名の異表記，貨幣単位の違い，時間帯の違い（現地時間と日本時間）等があるため，同じ事物を指すこれらの表現を複数個回答リストに含めたような重複があると考えられる場合の扱いを決めなければいけない．

回答の質に関する問題　同じ事物を指す上記の表現の中には，フルネームと略称のように情報の質が異なるものがある．日付や場所の場合は「00年」「00年1月3日」，「日本」「千葉県浦安市」のように粒度（詳細度）の異なるバリエーションがある．これら表現の質の問題を扱い，評価に反映させる必要がある．加えて，表現の問題ではなく，回答自体（指示されている事物）の質が異なると思える場合がある．例えば，記事中で事実もしくは伝聞として述べられているが，誤報もしくは発表者側の誤りにより事実と異なる数値や日付，記事中では確定的な予定として述べられているがその後に変更となった日付等を正当な正解と同じように扱ってよいのかには疑問が残り，その質の差に見合った評価が求められる．

列挙の体系の問題　可能な正解すべてを列挙するといっても，その列挙の体系が複数ある場合がある．「東海三県」と「三重県」「愛知県」「岐阜県」のように（一定の知識を前提とすれば）同じ情報が違う形で伝えられる場合がある．例示を含んだ「川魚，エビ，カニ等の魚介類」において．「川魚」「エビ」「カニ」「魚介類」は明らかに並べられるものではな

いが，「川魚」「エビ」「カニ」という列挙と「魚介類」という回答とのどちらが優れているかは自明ではない．この問題は粒度と関連して生じることが多い．あるイベントの開催地をそれが行われた国名で列挙するか都市名で列挙するかの選択もある．また，あるイベントが「12 月 10 日」と「12 月 20 日」の 2 回行われたとき，その開催日を「12 月」と答えてしまうと 2 回行われたという情報は伝わらない．この場合「12 月」と「12 月 10 日」のふたつを答えても，伝わる情報は「12 月 10 日」だけを答えた場合と同じである．表現の粒度の問題は表現の質の問題であるが，この例のようにその粒度が荒くなって他の回答と区別できなくなった時，そこにとどまらなくなる．加えて範囲表現等を正解範囲に含めているため，例えば，「8 世紀後期から 9 世紀初期」をひとつの要素とするリストと「8 世紀後期」「9 世紀初期」のふたつを要素とするリストとを等しく扱わなければならない．

これらの難しさを考慮し，可能な限り直観に合う評価を行うため，以下のような評価の枠組みを提案する．中心となるのは，正解セットという考え方の導入と回答の 2 種類の質を区別した多段階評価である．

各質問について，複数の正解セット CAS を用意する．ひとつの正解セットとは，ひとつの列挙の体系に対応するもので，上の例では，{「東海三県」} がひとつ，{「三重県」「愛知県」「岐阜県」} がひとつのセットをなす．また，{「12 月」} がひとつ，{「12 月 10 日」「12 月 20 日」} がひとつである．正解セット毎にそのセットの正解を網羅した際の係数 h $(0.0 < h \leq 1.0)$ が与えられる．多くの場合，その係数は 1.0 であるが，上例の {「12 月」} のセットの場合，このセットを網羅しても他方のセットの正解を網羅した場合の半分の情報しか与えられないとして，例えば係数 $h = 0.5$ が与えられる．

ある正解セットは，同じ事物を指す様々な正解表現 e の集まり（これを表現集合 ES と呼ぶ）の集まりである．人名における役職の有無や貨幣単位の違いのように同じ事物を指し，重複した回答として扱うべき表現に加えて，フルネームと略称のように情報を表現の質が異なるものや日付や場所において粒度が異なるものも，同じ事物を指す複数の表現として，ひとつの表現集合に含まれる正解表現となる．実際には正解判定は表現と根拠記事との対に対して行われるので，異なる根拠記事を持つ同じ表現も同じ表現集合に属するとして扱う．それぞれの表現集合についてそれが指すものの質に関する係数 g $(0.0 < g \leq 1.0)$ が付与される．表現集合中の正解表現それぞれには表現の質に関する係数 f $(0.0 < f \leq 1.0)$ が付与される．

システムが返した回答リスト O が与えられた時，ある正解セット CAS_i に関する精度 P_{CAS_i} と再現率 Q_{CAS_i} は図 2 の式で与えられる．これに基づいて F_{CAS_i} 値が求められ，最も大きい F_{CAS_i} 値を与える正解セット CAS_i を用いた評価がその回答リストに対する評価となる．なお，正解が存在しない質問については，回答数が 0 の場合（空リストを回答とした場合）に F 値

$$P_{CAS_i} = \frac{\sum_{ES \in CAS_i} \begin{cases} \max_{e \in O \cap ES} f(e) & \text{if } O \cap ES \neq \phi \\ 0 & \text{otherwise} \end{cases}}{|O| - |(O - \bigcup_{ES \in CAS_i} ES) \cap \bigcup_{ES' \in \bigcup_{j \neq i} CAS_j} ES'|}$$

$$R_{CAS_i} = \frac{h(CAS_i) * \sum_{ES \in CAS_i} g(ES) * \begin{cases} \max_{e \in O \cap ES} f(e) & \text{if } O \cap ES \neq \phi \\ 0 & \text{otherwise} \end{cases}}{\sum_{ES \in CAS_i} g(ES)}$$

$$F_{CAS_i} = \frac{2 * P_{CAS_i} * Q_{CAS_i}}{P_{CAS_i} + Q_{CAS_i}}$$

$$MF = \max_i F_{CAS_i}$$

図 2　評価尺度の *MF* 値の定義

1.0，それ以外は 0 とする．この定義による *F* 値を *MF* 値[6]，テストセットについてのその平均を *MMF* 値と呼ぶ．

この評価が意図しているのは，

- 表現の質は係数 f で表現し，質の低い表現を選んだ場合は精度再現率の分子となる正解数の当該部分にそれを乗じることでよりよい表現を回答した場合と差を付ける．
- 正解そのものの質は係数 g で表現し，再現率の分子分母の正解数両方にそれを乗じることで，再現率に正解の質を反映させる
- 同一物を指示する異表現はその同定をシステムの能力の一部と考え，同じ表現集合に属する正解を複数回答した場合は，その中で表現の質が最もよいものひとつを正解とし，それ以外は誤答として扱うことで精度を下げる．
- 正解の列挙については，ひとつの列挙の体系に基づいて回答することを期待し，それぞれの正解セットに従って採点を行い，最も高い評価となるセットの値を採用する．ただし，各セットでの採点において，そのセットでは誤答であるが，他のセットでの正解であるような回答は回答数に含めないことで，誤答と区別する．これにより様々な正解セットに含まれる正解を混在させた時の精度の減少を防ぎ，ペナルティをなくす．

一例として，「東京ディズニーランドはどこにありますか」という質問に「千葉県浦安市」「舞浜駅前」のふたつの正解があるとする．このふたつが同じ場所を指す異表現と考えるなら，同じ正解セットの同じ表現集合にこのふたつを含めることになる．その場合，一方を回答リストに含めればよく，両方を含めた場合，精度が下がる．これらふたつは違う情報であり，両方を列挙すべきであると判断した場合は，同じ正解セットの異なる表現集合に含める．この場合，両

[6] 若干の修正を含んでいるということで Modified の *M* を付けた．

方を回答リストに含めないと再現率が下がる．このふたつは異なる回答の仕方でありどちらもひとつで充分な情報を持っているとの判断であれば，これらふたつを異なる正解セットとする．この場合，一方を回答すればよく，両方を含めても精度は下がらない．両方回答すべき（同じ場所の別表現ではない）であるが，「千葉県浦安市」の方がより適切とする場合は，「舞浜駅前」の正解そのものの質に関する係数 g を落とす．この場合，例えば「千葉県浦安市」だけで再現率 0.67，「舞浜駅前」のみで 0.33 というような重み付けが可能となる．更に「千葉県」も正解とするが，これは「千葉県浦安市」と同じ場所を指し，表現として劣ると判断するのであれば，「千葉県浦安市」と同じ表現集合に含め，その表現に関する係数 f を落とせばよい．

4 参照用テストセット

一問一答形式の質問応答システム，特にリスト型課題にまだ研究の余地がある現状においては，システムの能力は様々な要因に左右され，情報アクセス対話における質問応答の能力だけでは決まらない．例えば，ある質問の正答率が低い時にその難しさがその文脈処理の側面にあるのかどうかは明らかでない．情報アクセス対話における質問応答でのシステム全体の能力を測定することがIADタスクの目的であるが，その改善に向けた分析が可能となるような材料が収集できることも望まれる．

そのような情報を得るための道具立てとして，あるテストセット（本節では主テストセットと呼ぶ）を用いたIADタスクの実施と並行して，その主テストセットから作成される2種類の参照用テストセットを用いて同じタスクを実施することを提案する．第一の参照用テストセットは，主テストセットに含まれる照応表現をすべて人手で解消し，それを補った独立の質問からなるセットである．第二の参照用テストセットは，主テストセットに含まれる照応表現のうち，代名詞＋助詞や連体詞等，表層に現れているものをすべて機械的に除去した独立の質問からなるセットである．こちらは意味的には，大半の質問が誰のものかを指定しないで誕生日を訊ねるような特定化が不充分なものとなるが，日本語であることが幸いして統語的には文法的である．図1に示した質問シリーズ series 2-20 に対応するこれら参照用テストセットの部分を図3に示す．第一の参照用テストセットの結果は文脈処理の上限，第二の結果は文脈処理なしで回答できる下限を示している．もちろん，文脈処理の結果得られる表現はひとつではないし，文脈処理が悪い影響を与えることも多いので，これらの結果は参考にとどまるが，このような参照用テストセットは技術の特徴を検討するのに有益である．

参照用テストセットによる実施が貴重な情報を提供する例として，NTCIR-4での実施での例を挙げる．表4は主テストセットのシリーズ最初の問題と2番目以降の問題について，上位10システムの *MMF* 値を平均したものと，第一の参照用テストセットについて，それに対応する値とを比較したものである．主テストセットでは当然，2番目以降の問題の平均 *MMF* 値が大

Series 2-20 に対応する第一の参照用テストセットの部分
ジョージ・マロリーはどこの国で生まれましたか。
ジョージ・マロリーの有名な言葉は何ですか。
ジョージ・マロリーが「そこに山があるからだ」と言ったのはいつのことですか。
ジョージ・マロリーが初めて山に登ったのは何歳の時ですか。
ジョージ・マロリーがエベレストの頂上付近で行方を絶ったのは何次遠征のときですか。
ジョージ・マロリーがエベレストの頂上付近で行方を絶ったのは何年のことですか。
ジョージ・マロリーが最後に目撃されたのはエベレストの何メートル付近ですか。
ジョージ・マロリーの遺体を発見したのは誰ですか。
Series 2-20 に対応する第二の参照用テストセットの部分
ジョージ・マロリーはどこの国で生まれましたか。
有名な言葉は何ですか。
言ったのはいつのことですか。
初めて山に登ったのは何歳の時ですか。
エベレストの頂上付近で行方を絶ったのは何次遠征のときですか。
何年のことですか。
最後に目撃されたのはエベレストの何メートル付近ですか。
遺体を発見したのは誰ですか。

図 3　参照用テストセットを構成する質問の例

表 4　質問の位置による評価（平均 *MMF* 値）の差

テストセット	シリーズ先頭	2 番目以降	全部
主	0.31	0.13	0.15
参照用 1	0.31	0.21	0.23

表 5　シリーズの型の違いによる評価（平均 *MMF* 値）の差

テストセット	狭義の収集	その他の収集	ブラウジング
主	0.20	0.16	0.13
参照用 1	0.23	0.19	0.30

きく落ちているが，参照用テストセットでもそれに対応する問題で平均 *MMF* 値が低くなっている．予想される理由は，あるトピックに関する一連の質問を行うと比較的簡単なものが先頭に来ることである．このような分析により，単にシリーズ最初の問題と 2 番目以降の問題についての *MMF* 値を比較して，文脈処理の困難さを過度に主張するという間違った結論を避けることができる．ちなみに，2 番目以降の問題について，参照用テストセットと主テストセットの平均 *MMF* 値の差は有意であることから，文脈処理の不十分さが 2 番目以降の問題の成績を悪くしていることも確認されている．

同様にシリーズの種類毎にみた上位 10 システムの平均 *MMF* 値を表 5 に示す．シリーズをす

べての照応表現が最初の質問で導入されたトピックを参照するという厳しい意味での収集型である狭義の収集型，その他の収集型，ブラウジング型に分類したものである．主テストセットではブラウジング型の平均 *MMF* 値が低いが，参照用テストセットにおいてはそれに対応する質問群について最も高い値が得られている．参照用テストセットにおけるこの違いは，シリーズ先頭の質問が比較的容易なのと同じ理由で，ブラウジング型のシリーズに含まれる様々なトピックに関する個々の質問は比較的容易なものになっていることによるのであろう．比較的簡単な個々の質問もブラウジング型のシリーズとして組織化されると難度の高いものになるということで，ブラウジング型シリーズにおける文脈処理が困難であることを再確認することができる．

第 2 の参照用テストセットの用途のひとつは，それぞれの質問について，その正解を得るために本当に文脈処理が必要かの情報が得られることである．例えば，図 1 の Series 2-20 の第 7 問に対応するものの平均 *MF* 値は，ふたつの参照用テストセットの間で殆ど差がなく，その値は主テストセットでの値よりも高い．これは，エベレストで最後に目撃された人間はマロリーの他には多くない（いない）ために，キーワードとしてマロリーがなくても正解を求められるためと思われる．同様の例で，「豊田章一郎氏が会長を務めていた自動車会社はどこですか。」「そこが 97 年に発売したハイブリッド車は何という名前ですか。」と続くシリーズにおいて，第 2 の質問に対応するものの *MF* 値もふたつの参照用テストセットの間で殆ど差がないが，日本でこの年に発売されたハイブリット車は 1 車種のみであり，会社名による限定が必要ないためであった．これらの情報は背景となる知識源の内容と関連し，事前に問題を検査して得るのは難しいが，参照用テストセットによって容易に明らかにすることができる．

5 関連研究

本稿での提案と最も近い取り組みは，TREC 2001 で行われた Context Task で，これは質問応答システムの文脈追跡（文脈処理）能力を測定するために一連の質問に回答させるというもので，基本的な目的は本稿の提案と同じである (Voorhees 2001)．このタスクの実施では，システムがある質問に正解できるかがそれ以前の質問に正解したかに依存しないという「予想に反する」結果が得られている．これは最初の質問によってそのシリーズの質問すべての回答を含んだ少数の記事が同定されてしまい，その後の質問に正解できるかは文脈処理の能力よりも特定のタイプの質問に回答できるかに依存してしまうためであるとされている．このため，このようなタスクは現状では文脈処理能力を測定するのに不適切と判断され，その後の TREC では実施されていない．

このような結果となったひとつの理由は，シリーズを構成する質問の数が 3 から 4 と少ないことにあると思われる．IAD タスクではひとつのシリーズは 7 つ程度の質問で構成することを

考えている．また，IADタスクでいうところのブラウジング型を含んでいないことも大きな原因であろう．TRECのContext Taskについては，隣り合う質問の回答のうち85%が同じパラグラフに存在したという報告 (Harabagiu, Moldovan, et al. 2001) があるが，NTCIR-4で用いたテストセットでは，隣り合う質問の少なくともひとつの回答が同じ記事（一概に比較できないが段落より大きい単位と言ってよいと考える）内に存在する割合は，収集型でも83%であったが，ブラウジング型では66%であった．シリーズ全体を考えれば，ブラウジング型の場合，ニューヨーク・ヤンキーズからキャンベルスープまでを含んだ記事はありえないので，最初の質問に関する処理だけでその後の質問に正解できる記事が得られることはありえない．収集型についても，すべてが狭義の収集型ではないので，そのトピックに関する記事すべてを検索してもそこから正しく回答を選択することは，何らかの文脈処理なしでは困難である．狭義の収集型についても，例えば，「小沢征爾」をキーワードとする記事は知識源中に155件あり，そのうちの22件が彼のウィーンフィルへの移籍を扱っているが，その中で彼の誕生日に言及しているものは2件のみである．また，収集型については，確かにある質問に回答できることと以前の質問への正解率との関係は不明確であるが，狭義の収集型であれば，そこに関係のある必然性はないし，そのことが文脈処理の不必要性の議論につながるとは思えない．加えて重要なことは，このようなタスク設計がレポート作成を目的とした情報アクセス対話という場面設定の状況に近いということであり，そこに現れる状況に対処する技術として必要とされている点である．

評価尺度についての*MF*値の提案は，IADタスクに限定されるものではなく，リスト型課題に共通するものである．TRECのQA Trackでも，2003年より正解数を指定しないリスト型課題が開始されている (Voorhees 2003)．評価には単純なF値が用いられている．2003年のこの課題の質問は37問とあまり多くなく，“List the names of chewing gums.”, “Who are female boxers?” 等，すべてが事物の列挙を求めるもので，その殆どは，“What Chinese provinces have a McDonald's restaurant?” のように回答のクラスが巧みに指定されており，粒度の問題が生じるような表現，例えば “Where in China does McDnald have a restaurant?” は避けられている．質問文のみからの判断であるが，問題が出る可能性のあるのはわずかに “What foods can cause allergic reaction in people?” の1問だけである．TRECにしてこのような状況であり，本稿で議論したようなリスト型課題の問題に注目した提案は著者の知る限り全く行われていない．

参照用のテストセットという考えについては，これもTREC-9において，同じ正解を意図した表現の異なる質問を多数テストセットに含めるという試みがなされている (Voorhees 2004)．参照用のテストセットという明確な考えはなく，そこから何が得られたかも明らかにされていないが，より深い分析のための情報を得る試みであったと思われる．この試みはその後続けられていない．一問一答型の質問応答システムも質問解析，文書選択，回答抽出等の複数のモジュールから構成されることを考えると，本稿で提案した参照用テストセットだけで充分な情報が得られるわけではないが，少なくとも情報アクセス対話のための質問応答技術をある程度まで区

別する役割を果たしていると考える．

対話的な質問応答システムの評価ということでは，テストセットの枠組みに基づかない，より実際に近い状況での実験の報告がある (Liddy, Diekema, and Yilmazel 2004; Kelly, Kantor, Morse, et al. 2006)．これらの実験と本稿で提案したテストセットによる評価は，情報検索技術の評価における検索実験での，現実状況での検証と研究室での検証 (岸田 1998) とにそれぞれ対応すると考えられる．前者は実際の利用場面により近い環境での評価となり，多種多様な情報が得られるが，それらの情報は複雑かつ非定型で分析も難しく，実験の実施も一般に高価である．一方で後者は，本来の利用場面の複雑さを切り捨て，理想化単純化された状況での能力を測定することになるが，得られるデータの相互比較が比較的容易で，テストセットの再利用が可能なこと等，その実施も安価である．このように，これらにはそれぞれの長所短所があり，相補的な役割を持っていると考えている．

6 おわりに

あるトピックに関して一連の情報アクセスを対話的に行うという状況で用いられる質問応答システムの能力を定量的に評価するためのタスク，IAD タスクを提案した．対話的な情報アクセスを模擬した実験を通じて，数量等を含む名称を正解の範囲とするような質問応答システムがそのような状況で有効であること，そのようなシステムは様々な照応表現を処理できる必要があることを示し，タスクが評価する技術の重要性を示唆した．IAD タスクは，対話的情報アクセスを対象として，そこで必要な質問応答技術が効果的に評価できるというその枠組みの独自性に加えて，質問中の参照表現を人手で解消もしくは機械的に削除した参照用テストセットを併用することで，情報アクセス対話におけるシステムの文脈処理能力をある程度まで切り離して評価できる枠組みを持っている．評価尺度についても自然な質問への応答を考えた場合に問題になる事例に配慮して，回答の列挙に複数の体系を許し回答の 2 種類の質を考慮に入れた多段階評価手法という，リスト型課題一般の評価手法に関する新しい提案を含んでいる．

付録

提案する IAD タスクが最先端の質問応答技術にとって，決して不可能な課題ではなく，同時に既に解決された課題でもないことを示すために，NTCIR-5 における QAC3 での実施において，高い評価を得た 3 チーム，7 システムについてその評価を示す．この実施では「施工ミス」「送電線切断」「墜落炎上」のような事象の複合名詞表現を正解範囲に含んでいたが，その位置づけが不明確なことから，今回の提案ではそれを除いている．その点を除けば，この実施は，本稿で提案している IAD タスクであり，事象の複合名詞表現を正解とする質問は少数であるため，

全体の傾向への影響は少ない．図4は，テストセット全体，各シリーズの先頭質問，2番目以降の質問について，*MMF* 値を示したものである．図5は，シリーズを収集型とブラウジング型に分類して，テストセット全体とそれらの *MMF* 値を比較している．これらのシステムに用いられている技術については，NTCIR-5 での QAC3 実施に関する報告 (Kato et al. 2004a; 加藤他 2006) に加え，(村田，内山，白土，井佐原 2007; Akiba 2006; Mori, Kawaguchi, and Ishioroshi

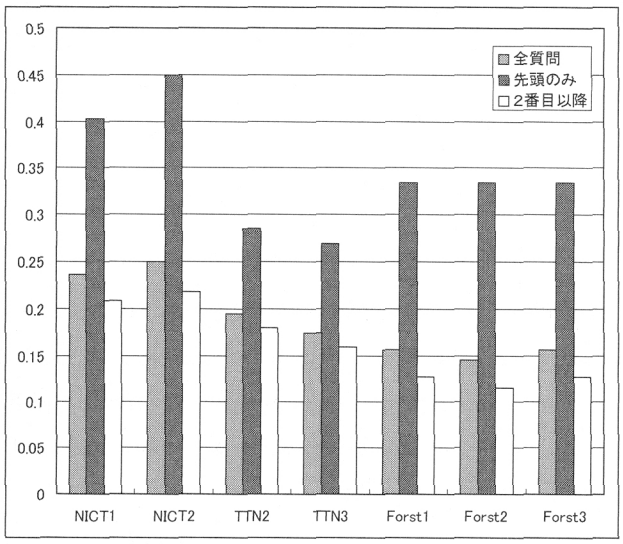

図 4　*MMF* 値による評価

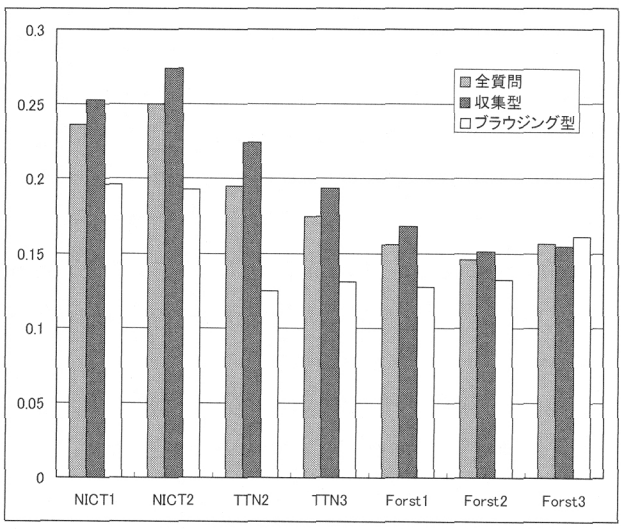

図 5　シリーズの型による *MMF* 値の差異

2007) に詳しい.

謝 辞

NTCIR-4 の QAC2 Subtask 3, NTCIR-5 の QAC3 に参加していただき, 貴重なコメントいただきました皆様に感謝します. 加えて, qac-j のメイリングリストでの議論に積極的に加わってくださった皆様にも感謝します. また, 本稿の中に直接活かすことはできませんでしたが, 村田真樹, 秋葉友良, 森辰則の 3 氏は, NTCIR-4 でのテストセットを用いた再度の実施を快く引き受けてくださいました. 御尽力にお礼申し上げます. 本研究の一部は, 国立情報学研究所との共同研究として支援されています.

参考文献

Akiba, T. (2006). "Exploiting Dynamic Passage Retrieval for Spoken Question Recognition and Context Processing towards Speech-driven Information Access." In *Proceedings of The International Conference on Language Resources and Evaluation (LREC)*, pp. 1530–1535.

ARDA. "AQUAINT Home Page: Advanced Question & Answering for Intelligence.", `http://www.ic-arda.org/InfoExploit/aquaint/`.

Burger, J., Cardie, C., et. al. "Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A).", `http://www-nlpir.nist.gov/projrcts/duc/roadmapping.html`.

Fukumoto, J., Kato, T., and Masui, F. (2003). "Question Answering Challenge (QAC-1) An Evaluation of Question Answering Tasks at the NTCIR Workshop 3." In *AAAI 2003 Spring Symposium New Directions in Question Answering*, pp. 122–133.

Harabagiu, S., Moldovan, D., et. al (2001). "Answering complex, list and context questions with LCC's Question-Answering Server." In *Proceedings of TREC 2001.*

Hickl, A., Lehmann, J., Williams, J., and Harabagiu, S. (2004). "Experiments with Interactive Question Answering in Complex Scenarios." In *Proceedings of HLT-NAACL2004 Workshop on Pragmatics of Question Answering*, pp. 60–69.

加藤恒昭, 福本淳一, 桝井文人, 神門典子 (2004a). "リスト型質問応答の特徴付けと評価指標." 情報処理学会自然言語処理研究会 2004-NL-163, pp. 115–112.

加藤恒昭, 福本淳一, 桝井文人, 神門典子 (2004b). "質問応答技術は情報アクセス対話を実現できるか." 情報処理学会自然言語処理研究会 2004-NL-162, pp. 145–150.

加藤恒昭, 福本淳一, 桝井文人, 神門典子 (2006). "情報アクセス対話に向けた質問応答技術の評

価ふたたび—NTCIR-5 QAC3での試み—.” 情報処理学会自然言語処理研究会 2004-NL-172, pp. 55–62.

Kato, T., Fukumoto, J., and Masui, F. (2004a). “An Overview of NTCIR-5 QAC3.” In *Proceedings of Fifth NTCIR Workshop Meeting*, pp. 361–372.

Kato, T., Fukumoto, J., and Masui, F. (2004b). “Question Answering Challenge for Information Access Dialogue—Overview of NTCIR4 QAC2 Subtask3—.” In *Working notes on the Fourth NTCIR Workshop Meeting*, pp. 291–296.

Kato, T., Fukumoto, J., Masui, F., and Kando, N. (2005). “Are Open-domain Question Answering Technologies Useful for Information Access Dialogues? —An empirical study and a proposal of a novel challenge—.” *ACM TALIP (Trans. of Asian Language Information Processing)*, **4** (3), pp. 243–262.

Kato, T., Fukumoto, J., Masui, F., and Kando, N. (2006). “WoZ Simulation of Interactive Question Answering.” In *Proceedings of HLT-NAACL2006 Workshop on Interactive Question Answering*, pp. 9–16.

Kelly, D., Kantor, P., Morse, E., et. al (2006). “User-Centered Evaluation of Interactive Question Answering Systems.” In *Proceedings of HLT-NAACL2006 Workshop on Interactive Question Answering*, pp. 49–56.

岸田和明 (1998). 情報検索の理論と技術. 勁草書房.

Liddy, E. D., Diekema, A. R., and Yilmazel, O. (2004). “Context-Based Question-Answering Evaluation.” In *Proceedings of the 27th Annual International ACM SIGIR Conference*, pp. 508–509.

Mani, I., House, D., et. al (1998). “The TIPSER SUMMAC text summarization evaluation final report.” Tech. rep. MTR98W0000138, The MITRE Corporation.

Mori, T., Kawaguchi, S., and Ishioroshi, M. (2007). “Answering Contextual Questions Based on the Cohesion with Knowledge.” *International Journal of Computer Processing of Oriental Languages (IJCPOL)*, **20** (2&3), pp. 115–135.

NII. “NTCIR (NII-NACSIS Test Collection for IR Systems) Project Home Page.”, `http://research.nii.ac.jp/ntcir/index-ja.html`.

NIST. “TREC Home Page.”, `http://trec.nist.gov/`.

小川泰嗣，佐々木裕，増山繁，村田真樹，吉岡真治 (2002). “参加者から見たNTCIR.” 人工知能学会誌, **17** (3), pp. 306–311.

関根聡，影浦峡，奥村学，乾健太郎 (2005). “研究の場としての評価型ワークショップになるために.” 言語処理学会第11回年次大会併設ワークショップ「評価型ワークショップを考える」.

Small, S., Shimizu, N., et. al (2003). “HITIQA: A Data Driven Approach to Interactive Question

Answering: A Preliminary Report." In *AAAI 2003 Spring Symposium New Directions in Question Answering*, pp. 94–104.

Voorhees, E. M. (2001). "Overview of the TREC 2001 Question Answering Track." In *Proceedings of TREC 2001*.

Voorhees, E. M. (2003). "Overview of the TREC 2003 Question Answering Track." In *Proceedings of TREC 2003*, pp. 14–27.

Voorhees, E. M. (2004). "Overview of the TREC 2004 Question Answering Track." In *Proceedings of TREC 2004*.

Voorhees, E. M. (2005). "Question Answering in TREC." In Voorhees, E. M. and Harman, D. K. (Eds.), *TREC Experiment and Evaluation in Information Retrieval*. The MIT Press.

Voorhees, E. M. and Tice, D. M. (2000). "Building a Question Answering Test Collection." In *Proceedings of the 23rd Annual International ACM SIGIR Conference*, pp. 200–207.

村田真樹，内山将夫，白土保，井佐原均 (2007). "シリーズ型質問文に対して単純結合法を利用した逓減的加点質問応答システム." システム制御情報学会論文誌, **20** (8), pp. 338–346.

略歴

加藤　恒昭：1983 年東京工業大学大学院総合理工学研究科修士課程修了．同年，日本電信電話公社（現 NTT）に入社．2000 年東京大学大学院総合文化研究科言語情報科学専攻准教授，現在に至る．1993 年米国ロチェスター大学客員研究員．2005 年米国 USC/ISI 客員研究員．博士（工学）．質問応答，情報編纂，語彙意味論に関する研究に従事．

福本　淳一：1986 年広島大学大学院工学研究科博士前期課程修了．同年，沖電気工業（株）に入社．1992–94 年英国マンチェスター科学技術大学 Ph. D. コース．2000 年立命館大理工学部助教授．2004 年米国 USC/ISI 客員研究員．2006 年立命館大情報理工学部教授，現在に至る．Ph. D．質問応答システム，情報抽出，談話構造解析の研究に従事．

桝井　文人：1990 年岡山大学理学部・地学科卒業．同年，沖電気工業（株）に入社．2000 年三重大学工学部情報工学助手．2004–05 年北海道大学情報科学研究科研究員．現在，三重大学大学院工学研究科助教．博士（工学）．自然言語処理，教育工学，設備保全などの研究に興味を持つ．

神門　典子：1994 年慶応義塾大学文学研究科博士課程修了．博士（図書館・情報学）．同年学術情報センター助手．米国シラキウス大学情報学部，デンマーク王立図書館情報大学客員研究員を経て，1998 年学術情報センター助教授．2000 年国立情報学研究所助教授．2004 年同教授．現在に至る．テキスト構造

を用いた検索と情報活用支援，言語横断検索，情報検索システムの評価等の研究に従事.

(2007 年 9 月 3 日　受付)
(2008 年 1 月 11 日　再受付)
(2008 年 3 月 3 日　採録)